Exercise 1: Splitting criteria

Given are the data set

x	1.0	2.0	7.0	10.0	20.0
y	1.0	1.0	0.5	10.0	11.0

and the same with log-transformed feature x:

$\log x$	0.0	0.7	1.9	2.3	3.0
y	1.0	1.0	0.5	10.0	11.0

a) Compute the first split point the CART algorithm would find for each data set (with pen and paper or in `R`, resp. `Python`).

b) State the optimal constant predictor for a node $\mathcal{N}$ when minimizing the empirical risk under $L2$ loss and explain why this is equivalent to minimizing "variance impurity".

Exercise 2: Impurity reduction

The fractions of the classes $k = 1, \dots, g$ in node $\mathcal{N}$ of a decision tree are $\pi_1^{(\mathcal{N})}, \dots, \pi_g^{(\mathcal{N})}$. Assume we replace the classification rule in node $\mathcal{N}$

$$\hat{k} \mid \mathcal{N} = \arg\max_k \pi_k^{(\mathcal{N})}$$

with a randomizing rule

$$\hat{k} \sim \text{Cat}\left(\pi_1^{(\mathcal{N})}, \dots, \pi_g^{(\mathcal{N})}\right),$$

in which we draw the classes in one node from the categorical distribution of their estimated probabilities (i.e., class k is predicted with probability $\pi_k^{(\mathcal{N})}$).

Compute the expected MCE in node $\mathcal{N}$ for data distributed i.i.d. like the training data. What do you notice?

(*Hint*: The observations and the predictions using the randomizing rule follow the same distribution.)